Published in: Hardie, Andrew/Love, Robbie (eds.): Corpus Linguistics 2013. Abstract Book. - Lancaster: UCREL, 2013. S. 23-25.

Robust corpus architecture: a new look at virtual collections and data access

Piotr Bański
IDS Mannheim / University of Warsaw
banski@ids-mannheim.de

Elena Frick
IDS Mannheim
frick@ids-mannheim.de

Michael Hanl
IDS Mannheim
hanl@ids-mannheim.de

Marc Kupietz
IDS Mannheim
kupietz@ids-mannheim.de

Carsten Schnober
IDS Mannheim
schnober@ids-mannheim.de

Andreas Witt
IDS Mannheim
witt@ids-mannheim.de

1 Introduction

The present contribution has two logical components: the first presents the basic aims and design principles of KorAP – a new corpus analysis platform that is being developed at the Institut für Deutsche Sprache in Mannheim. In the second part, we concentrate on two closely related issues that have arisen in the process of the development of the internal data architecture for KorAP but have consequences for innovative corpus design in general. These issues prompt a reformulation of the definition of the concept of virtual collections and a new assessment of the consequences of this reformulated concept for the practical considerations of access permissions and security in general.

2 Aims

KorAP (Korpusanalyseplattform der nächsten Generation, cf. Bański et al. 2012), currently in the prototype phase, is an innovative corpus analysis platform designed to address the demands of modern linguistic research. The platform is intended to facilitate new linguistic findings by making it possible to manage and analyse primary data and annotations in, eventually, the petabyte range, while at the same time fully satisfying the demands of a scientific tool, by both allowing an undistorted view of the primary linguistic data, and giving equal status to the various possible analyses of those data: in KorAP documents, the primary ("raw") text is physically separated from all its possible interpretations, from which the user may choose and which she may compare (this is an example of radical stand-off architecture, similar to that used in the American National Corpus, cf. Ide and Suderman, 2006).

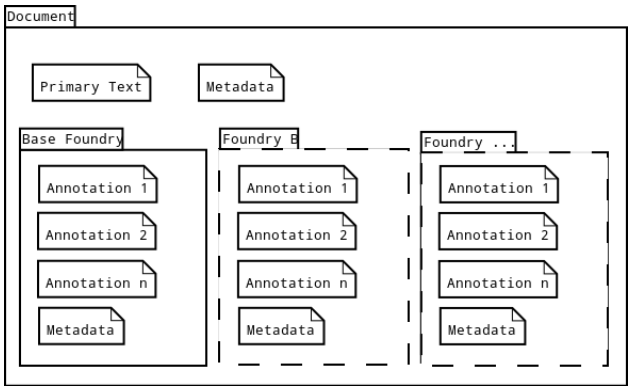

Figure 1. KorAP data model, where the primary ("raw") text is separated from annotations, organized into "foundries".

An additional important aim of the project is to make corpus data as openly accessible as possible in light of unavoidable legal restrictions, for instance by providing a sandbox that enables users to apply their own tools, working on data that cannot be released, or by supporting distributed virtual collections (Kupietz et al. 2010). The KorAP software itself will be released under an open license.

3 Virtual collections and security

The term "virtual collections" was first introduced, i.e. imported from the context of digital libraries, by the D-SPIN project (the former name of the German part of CLARIN, cf. Bankhardt 2009) as a generalization of "virtual corpora" and as a fundamental concept for the development of a standardized way to persistently identify research data consisting of language resources, in order to facilitate the implementation of maxims such as replicability, in linguistics and adjacent disciplines.

It has been further elaborated within the CLARIN project, also in the form of a basic implementation of a registry for virtual collections ("CLARIN-VCR")[1], and has been characterized as "distributed collections of corpora or data" (Broeder et al. 2007) that can "include a large number of resources created by different teams at different institutions" (ISO 24619:2011). An illustration of one practical implementation of the concept is provided by DeReKo (Deutsches Referenzkorpus, Kupietz et al. 2010), cf. Figure 2. Whenever applied to corpora, this term has usually denoted some amount of text accompanied by a single instance of grammatical description, most often inline – embedded in the text by means of XML elements or attributes.

[1] See http://clarin.ids-mannheim.de/vcr/app/public

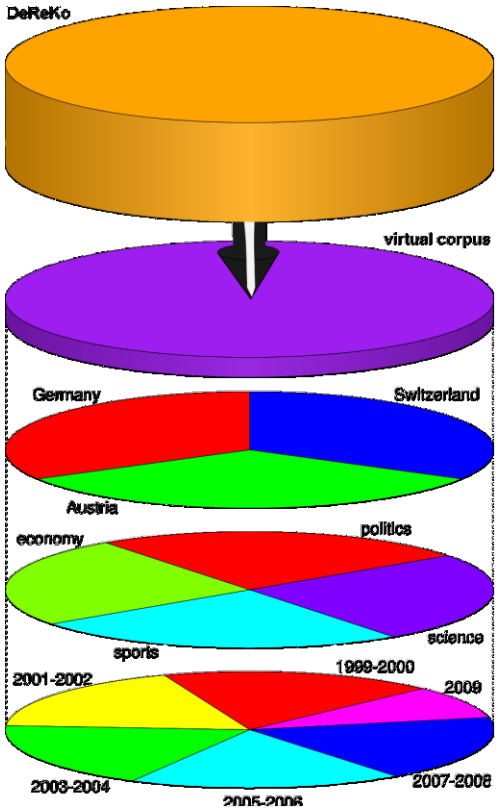

Figure 2. Illustration of DeReKo (Deutsches Referenzkorpus) seen as a "primordial sample", from which individual virtual collections can be created according to metadata-based choices, in this case: country of origin, topic, and time period. Copied from Kupietz et al. (2010).

This by now widely adopted characterisation of virtual collections and the current state of the CLARIN-VCR have proven insufficient for a tool such as KorAP, which explicitly allows for many concurrent annotation layers to provide equally valid points of view on the underlying data. KorAP explicitly distinguishes the primary ("raw") text of the individual corpus documents from its linguistic descriptions, even down to the level of supporting multiple tokenizations, with hierarchies of annotation layers built upon each tokenization stream.

KorAP's robust support for metadata describing not only the primary text but also each annotation layer, together with the traditional concept of virtual collections, result in multi-faceted virtual collections that, apart from texts, can combine e.g. annotations produced by the same tool or within the same school of linguistic thought, including collections of resources bearing the same distribution licenses. This is illustrated in Fig. 3.

As the complexity of linguistic resources increases, assigning permissions directly to users becomes an administrative challenge and as a consequence poses high security risks. To deal with such security issues and improve administration and performance, existing corpus analysis systems and modern research platforms such as TextGrid (TextGrid-Konsortium 2009: 18f) opt most often for role- and/or group-based approaches (MPI 2006: 83f.). Although these concepts offer graded access control by allowing grouping to corpora and virtual collections, they lack flexibility to handle KorAP's demands for access control. KorAP's ability to separate the raw text from its various grammatical representation results in a large variety of collections that can be created.

Furthermore, KorAP has been designed to support

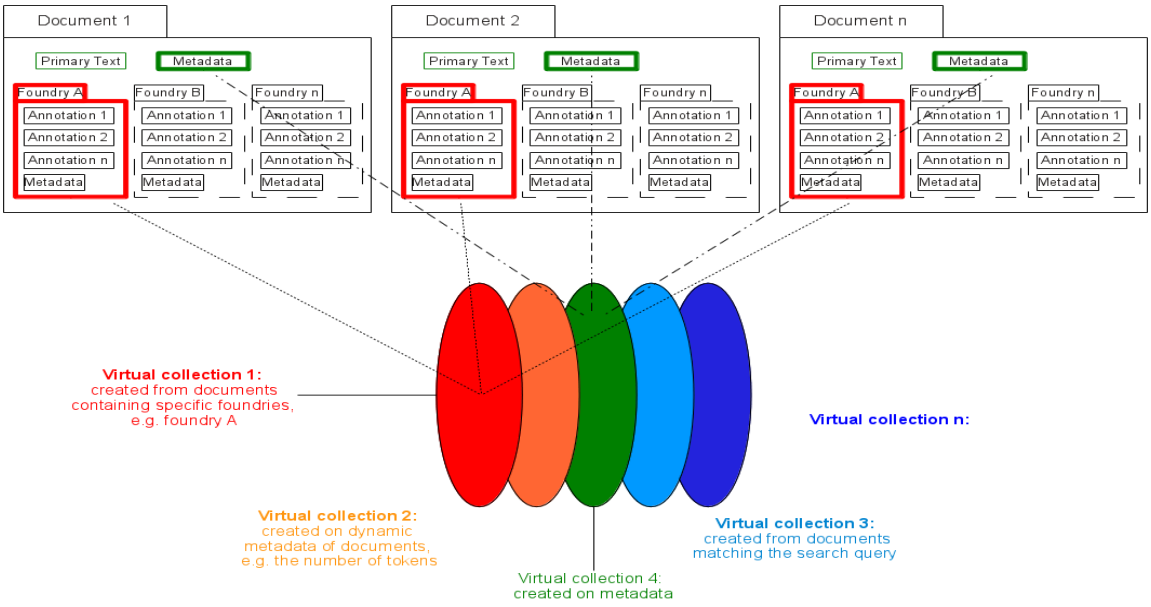

Figure 3. A sample of ways in which KorAP virtual collections can be created

user-supplied corpora, which imposes additional requirements on an access control system: reliance on mappings to roles does not allow for the envisioned definition of fine-grained access control policies, and a group-based approach would result in massive overkill.

References

Bankhardt (2009): D-Spin – Eine Infrastruktur für deutsche Sprachressourcen. In: Sprachreport 1/2009. S. 30-31 – Mannheim: Institut für Deutsche Sprache, 2009. (Sprachreport 1/2009)

Bański, P. Fischer, P.M., Frick, E., Ketzan, E., Kupietz, M., Schnober, C., Schonefeld, C., and Witt, A. 2012. The new IDS corpus analysis platform: Challenges and prospects. In *Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul.

Broeder, D., Declerck, Th., Kemps-Snijders, M., Keibel, H., Kupietz, M., Lemnitzer, L., Witt, A., Wittenburg, P. (2007): Citation of Electronic Resources: proposal for a new work item in ISO TC37/SC4. ISO TC37/SC4-Dokument N366.

Ide, N., Suderman, K. (2006). Integrating Linguistic Resources: The American National Corpus Model. In Proceedings of the Fifth Language Resources and Evaluation Conference (LREC), Genoa, Italy.

ISO 24619:2011. Language Resource Management – Persistent Identification and sustainable Access in Language Technology Applications. (PISA). Technical report. International Organization for Standardization.

Kupietz, M/Belica, C/Keibel, H/Witt, | (2010): The German Reference Corpus DeReKo: A primordial sample for linguistic research. In: Calzolari, Nicoletta et al. (eds.): Proceedings of the seventh conference on International Language Resources and Evaluation (LREC 2010), S. 1848-1854. http://www.lrec-conf.org/proceedings/lrec2010/pdf/414_Paper.pdf

Kupietz, M/Bankhard, C (eds.) (2009): D-SPIN Report R7.1 – Legal Aspects in the Provision of Language Resources: The German Context. (http://weblicht.sfs.uni-tuebingen.de/Reports/D-SPIN_R7.1.pdf)

Kupietz, M/Bankhard, C (eds.) (2010): D-SPIN Report R7.3 – Initial Localisation of CLARIN Best Practices and Business Models. (http://weblicht.sfs.uni-tuebingen.de/Reports/D-SPIN_R7.3.pdf)

Max-Planck-Institute for Psycholinguistics (MPI). (2006). DAM-LR Distributed Access Management for Language Resources – Deliverable 8.1 Definition Report (pp. 1–92). Nijmegen. (http://www.mpi.nl/dam-lr/documents.html)

TextGrid-Konsortium (2009): Abschlussbericht – Öffentliche Fassung. http://www.textgrid.de/fileadmin/berichte-1/abschlussbericht-1.pdf